Journal of Multivariate Analysis **63**, 242–258 (1997)
Article No. MV971704

A New Class of Consistent Estimators for Stochastic Linear Regressive Models*

Hong-Zhi An

Chinese Academy of Sciences, Beijing, China

Fred J. Hickernell

Hong Kong Baptist University, Hong Kong, China

and

Chinese Academy of Sciences, Beijing, China

In this paper we propose a new approach for estimating the unknown parameter in the stochastic linear regressive model with stationary ergodic sequence of covariates. Under mild conditions on the joint distribution of the covariate and the error, the estimator constructed is shown to be strongly consistent in two important special cases: (1) The sequence of (variate, covariate) is independent identically distributed (i.i.d.), and (2) the sequence of variates is a stationary autoregressive series. The asymptotical normality is also discussed under more assumptions on the distribution of the covariate. © 1997 Academic Press

1. INTRODUCTION

Linear regression of a random variable on another has been widely used in practice. The linear regression of random variable y on random vector x can be described by the following model

$$y = b^{\tau}x + \varepsilon \tag{1.1}$$

Received September 15, 1993; revised June 6, 1997.

Key words and phrases: Asymptotic normality, autoregressive model, consistent estimator, robustness, stochastic regressive model.

* This research was by the NSF of China and by a Hong Kong UPGC-RGC grant.

† To whom correspondence should be addressed.

0047-259X/97 $25.00

For the samples $(y_1, x_1^\tau), ..., (y_n, x_n^\tau)$ from (y, x^τ), the analogous model is realized,

$$y_t = b^\tau x_t + \varepsilon_t, \qquad t = 1, 2, ..., n. \tag{1.2}$$

Model (1.1) is called a stochastic regressive model because of the randomization of the x_i's (e.g., Lai and Wei, 1982).

Estimation of the unknown parameter b based on the sample (y_t, x_t^τ), $t = 1, ..., n$, is a crucial problem in the analysis of a stochastic regressive model (1.1). Various estimation techniques have been developed, such as the least squares method and the least absolute deviation method. The least squares estimator (LSE) of b is popular because of its computational simplicity and mathematical beauty, especially in the Gaussian case. The M-estimator of b such as the least absolute deviation estimator (LADE) has been the subject of much attention because of its robust properties. To ensure strong consistency of these estimators, some conditions on the distributions of x_t and ε_t must be imposed. For the LSE, the finite second moment of ε_t is needed (e.g., Lai and Wei, 1982), and for the LADE some conditions on the distribution of ε_t are also indispensable (e.g., An and Chen, 1983; Chen and Wu, 1988). However, in some cases, especially in the exploratory stage of data analysis, it is difficult to verify any such assumptions on the distribution of x_t and of ε_t. Hence we were motivated to seek a new procedure to obtain a strongly consistent estimator of b under less restrictive assumptions on the distributions of x_t and ε_t. Indeed, for two important cases discussed below, the assumptions required are the least restrictive in the sense that they correspond only to requiring that model (1.2) be well-defined.

There are two important cases that have motivated our analysis:

Independent and Identically Distributed (I.I.D.) Case. Let $(y_1, x_1^\tau), ..., (y_n, x_n^\tau)$ be i.i.d. samples from (y, x^τ) which satisfy model (1.1). Furthermore, let ε be independent of x in model (1.1).

Stationary Autoregressive Case. $x_t = (y_{t-1}, ..., y_{t-p})^\tau$, and $b = (b_1, ..., b_p)^\tau$ so that

$$y_t = b_1 y_{t-1} + \cdots + b_p y_{t-p} + \varepsilon_t, \tag{1.3}$$

where $\{\varepsilon_t\}$ is an i.i.d. sequence and ε_t is independent of $\{y_s : s < t\}$. Furthermore, let $b = (b_1, ..., b_p)^\tau$ satisfy the following stability condition:

$$1 - b_1 u - \cdots - b_p u^p \neq 0 \qquad \text{for} \quad |u| \leqslant 1 \tag{1.4}$$

and let

$$E[\max\{1, \log |\varepsilon_t|\}] < \infty \tag{1.5}$$

in order to guarantee the existence of a stationary solution of model (1.3) (An, 1990).

Below we give two sets of assumptions under which the new estimator of b will be derived and analyzed. These assumptions are the least restrictive in the sense that they correspond only to requiring that the model (1.2) be well-defined for the two cases given above.

Assumption A. (i) the series $\{(y_t, x_t^\tau)\}$ is a stationary and ergodic sequence with the same distribution as (y, x^τ) which satisfies (1.1);

(ii) x is non-degenerate, i.e., there is no nonzero vector c such that $c^\tau x$ is a degenerate random variable;

(iii) the series $\{\varepsilon_t\}$ is i.i.d., and ε_t is independent of $\{x_s : s \leqslant t\}$, hence ε is independent of x in model (1.1).

Assumption A(ii) is a necessary condition for model (1.2) to be well-defined. For the i.i.d. case specified above, Assumptions A(i) and A(iii) are satisfied naturally, and consequently Assumption A becomes the least restrictive possible.

Assumption B. (i) (1.4) and (1.5) hold;

(ii) ε_t is non-degenerate;

(iii) $\{\varepsilon_t\}$ is an i.i.d. series, and ε_t is independent of $\{y_s : s < t\}$.

For the stationary autoregressive case specified above, ε_t must be non-degenerate in order for (1.3) to be well-defined. Assumption B(ii) corresponds to Assumption A(ii) for the i.i.d. case. Thus Assumption B is the least restrictive assumption under which the stationary autoregressive model (1.3) is well-defined and has a stationary solution. In the literature of time series analysis, many strongly consistent theorems about the estimators of b have been established under some moment conditions (e.g. An *et al.*, 1982; An and Chen, 1983). Although in recent studies some new estimators of b for non-negative autoregressive models are shown to be strongly consistent without any moment conditions on ε_t (e.g., Anděl, 1989; An, 1990), these results assume ε_t must be positive in addition to Assumption B. In estimating b for model (1.3), again we must avoid estimating $E\varepsilon_t$. On the other hand, we should note that model (1.3) may be physically meaningful whether $E|\varepsilon_t| = \infty$ or not. Moreover, sometimes ε_t may be positive (e.g. Bell and Smith, 1986), and even $E\varepsilon_t = \infty$ occurs.

In this paper, we propose a new class of consistent estimator of the regressive parameter b for model (1.2) under the conditions assumed above on the distributions of x_t and ε_t. Section 2 describes the construction of our new estimator in detail and Section 3 provides some simulation results. Section 4 studies the strong consistency of the estimator for the above two

special cases of (1.2). Section 5 discusses the limiting distribution of the new estimator of b.

2. A NEW ESTIMATOR OF THE PARAMETER b

The LSE of b for model (1.2) is related to the residual sum of squares, an objective function, which is defined as

$$S_n(\beta) = \sum_{t=1}^{n} (y_t - \beta^\tau x_t)^2.$$

By minimizing $S_n(\beta)$, we obtain the LSE, while the LADE is derived based on the absolute deviation objective function. To construct a new estimator, we define a new criterion function first. By maximizing the new criterion function we can obtain a new estimator of b. Suppose that y, x, and ε satisfy model (1.1), i.e., suppose that

$$\varepsilon = y - b^\tau x. \tag{2.1}$$

Consider the vector parameter β in R^p, and let $\varphi(t, \beta)$ denote the characteristic function of variable $(y - \beta^\tau x)$, i.e.,

$$\varphi(s, \beta) = Ee^{is(y - \beta^\tau x)}. \tag{2.2}$$

Let

$$A(\beta) = \int |\varphi(x, \beta)|^2 \, w(t) \, dt, \tag{2.3}$$

where $w(\cdot)$ is a continuous density kernel function to be chosen under the conditions

$$w(s) = w(-s) \geqslant 0, \qquad \int |s| \, w(s) \, ds < \infty. \tag{2.4}$$

By Assumption A(iii) or B(iii), ε and x are independent, so (2.1) and (2.2) imply that

$$\begin{aligned} \varphi(s, \beta) &= Ee^{is(y - \beta^\tau x)} \\ &= Ee^{is(\varepsilon + (b - \beta)^\tau x)} \\ &= Ee^{is\varepsilon} Ee^{is(b - \beta)^\tau x} \\ &= \varphi_\varepsilon(s) \, \varphi_x((b - \beta)^\tau s), \end{aligned} \tag{2.5}$$

where $\varphi_\varepsilon(\cdot)$ and $\varphi_x(\cdot)$ are the characteristic functions of ε and x, hence ε_t and x_t as well, respectively. Because x is non-degenerate by Assumption A(ii) or B(ii), it is easy to see that if $\beta \neq b$,

$$|\varphi_x((b-\beta)^\tau s)| < 1, \tag{2.6}$$

except for countable many values of t. Combining (2.5) and (2.6) with (2.3) we know that

$$A(b) = \sup_{\beta \in R^p} A(\beta) > A(\alpha) \qquad \text{for every} \quad \alpha \neq b. \tag{2.7}$$

So far we have derived an objective function $A(\beta)$, depending on the distribution of (y, x^τ). In order to construct a criterion function which depends only on the samples (y_t, x_t^τ), $t = 1, ..., n$, $A(\beta)$ is written in another form.

Let $F_\beta(\cdot)$ denote the distribution of $(y - \beta^\tau x)$. By (2.2) and (2.3),

$$\begin{aligned} A(\beta) &= \int \left(\int e^{it(u-v)}\, dF_\beta(u)\, dF_\beta(v) \right) w(t)\, dt \\ &= \int \left(\int e^{it(u-v)} w(t)\, dt \right) dF_\beta(u)\, dF_\beta(v) \\ &= \int \varphi_w(u-v)\, dF_\beta(u)\, dF_\beta(v), \end{aligned} \tag{2.8}$$

where $\varphi_w(\cdot)$ is the characteristic function corresponding to the kernel $w(\cdot)$

$$\varphi_w(u-v) = \int e^{it(u-v)} w(t)\, dt. \tag{2.9}$$

For a given β, let $\hat{F}_\beta(\cdot)$ denote the empirical distribution of $(y - \beta^\tau x)$, based on the samples $y_t - \beta^\tau x_t$, $t = 1, ..., n$,

$$\hat{F}_\beta(u) = \frac{1}{n} \sum_{t=1}^{n} I(y_t - \beta^\tau x_t < u),$$

where $I(\cdot)$ denotes the indicator function. Finally, replacing $F_\beta(\cdot)$ by $\hat{F}_\beta(\cdot)$ in (2.8), we get

$$\begin{aligned} A_n(\beta) &= \int \varphi_w(u-v)\, d\hat{F}_\beta(u)\, d\hat{F}_\beta(v) = \frac{1}{n^2} \sum_{t=1}^{n} \sum_{s=1}^{n} \varphi_w(y_t - \beta^\tau x_t - y_s + \beta^\tau x_s) \\ &= \frac{1}{n^2} \sum_{t=1}^{n} \sum_{s=1}^{n} \varphi_w(y_t - y_s - \beta^\tau (x_t - x_s)), \end{aligned} \tag{2.10}$$

which is the criterion function proposed in this paper.

The desired estimator of the parameter b in model (1.2) is $\hat{b}_U$, the value that maximizes $A_n(\beta)$, i.e.,

$$A_n(\hat{b}_U) = \sup_{\beta \in R^p} A_n(\beta). \tag{2.11}$$

For model (1.3), the maximum value is taken over a subset D_p of R^p, which is defined by the stationarity condition, i.e.,

$$D_p = \{\beta = (\beta_1, ..., \beta_p)^\tau : 1 - \beta_1 u - \cdots - \beta_p u^p \neq 0, \text{ for } |u| \leqslant 1\}. \tag{2.12}$$

Now we consider the choice of the kernel function $w(\cdot)$ used in (2.10). Many density functions can be chosen as $w(\cdot)$, for example, the densities of the normal distribution $N(0, a^2)$, the uniform distribution on $(-a, a)$, the symmetric exponential distribution, i.e.,

$$w(s) = (a/2)\, e^{-a|s|}, \tag{2.13}$$

etc. Ideally, $w(\cdot)$ should have a simple form, and as a referee pointed out, it should have closed-form Fourier transforms to save computational time and the estimator should not be too sensitive to the parameters appearing in $w(\cdot)$. By taking account of the above principles, we prefer to use the density

$$w(s) = (a/3)(2e^{-a|s|} - e^{-2a|s|}), \tag{2.14}$$

where the choice of a will be discussed later on. In this case, it is easy to check that

$$\varphi_w(u) = 4a^4/(a^2 + u^2)(4a^2 + u^2), \tag{2.15}$$

and then

$$A_n(\beta) = (4a^4/n^2) \sum_{t=1}^{n} \sum_{j=1}^{n} \{[a^2 + (y_t - y_j - \beta^\tau(x_t - x_j))^2] \times [4a^2 + (y_t - y_j - \beta^\tau(x_t - x_j))^2]\}^{-1}. \tag{2.16}$$

$\hat{b}_U$ is taken to maximize $A_n(\beta)$ of (2.16). In the next section we report a small simulation study.

3. A SIMULATION STUDY

In order to compare the estimator $\hat{b}_U$ with the LSE $\hat{b}_L$ and the LADE b_M, some simulation results are given in this section.

We choose the following model to generate data, i.e.,

$$y_t = bx_t + \varepsilon_t, \qquad t = 1, ..., n, \tag{3.1}$$

where $b = 0.8$, $n = 100$, and the x_t are i.i.d. with common uniform distribution on $(0, 10)$, and ε_t are i.i.d. and independent of the x_t. The following four distributions of ε_t are separately investigated:

(1) $N(0, 9)$, the normal distribution with zero mean and variance 9,

(2) $C(0, 1)$, the Cauchy distribution with density $1/\pi(1 + x^2)$,

(3) the Bernoulli distribution $\Pr(\varepsilon = -3) = \Pr(\varepsilon = -3) = \frac{1}{2}$,

(4) an asymmetric distribution combining (2) and (3),

$$\Pr[\varepsilon \leqslant z] = \begin{cases} 0 & \text{for} \quad -\infty < z < -3 \\ \dfrac{1}{2} & \text{for} \quad -3 \leqslant z < 0 \\ \displaystyle\int_{-\infty}^{z} \frac{1}{\pi(1+x^2)}\, dx & \text{for} \quad 0 \leqslant z < +\infty. \end{cases}$$

In each case, 500 independent simulations of the series $(y_1, x_1), ..., (y_n, x_n)$ are performed, and in each simulation the estimators $\hat{b}_U$, $\hat{b}_L$, and $\hat{b}_M$ are calculated. Their medians are given in Table I. The numbers in parentheses are their empirical interquartile distances.

Here, $\hat{b}_U$ is obtained by the procedures mentioned above with the kernel function $w(\cdot)$ of (2.14). The tuning parameter a is set to be three different values, $a = 0.5$, 2, and $1/\sqrt{n}$ times the sample interquartile distance of the y_t. Thus in Table I there are three lines for the estimator $\hat{b}_U$.

In each simulation we calculate $\hat{b}_L$ in two ways, i.e., by solving

$$\sum_{t=1}^{n} (y_t - \hat{b}_L x_t)^2 = \inf_{\beta} \sum_{t=1}^{n} (y_t - \beta x_t)^2 \tag{3.2}$$

and

$$\sum_{t=1}^{n} (y_t - \hat{b}_0 - \hat{b}_L x_t)^2 = \inf_{\beta_0, \beta} \sum_{t=1}^{n} (y_t - \beta_0 - \beta x_t)^2 \tag{3.3}$$

separately. Thus Table I shows two values for the estimator $\hat{b}_L$.

Analogously, in Table I there are two estimators for $\hat{b}_M$ also, given by

$$\sum_{t=1}^{n} |y_t - \hat{b}_M x_t| = \inf_{\beta} \sum_{t=1}^{n} |y_t - \beta x_t| \tag{3.4}$$

TABLE I

Simulations for Model (3.1) with $b = 0.8$

Distribution of ε_t	$N(0, 9)$	$C(0, 1)$	*BERN*	*ASYM*
$\hat{b}_L$ (by (3.2))	0.7980 (0.1063)	0.7934 (0.6358)	0.7949 0.1290)	0.8059 (0.3418)
$\hat{b}_L$ (by (3.3))	0.7982 (0.0562)	0.7905 (0.3120)	0.8015 (0.0710)	0.8094 (0.2446)
$\hat{b}_M$ (by (3.4))	0.7960 (0.1674)	0.7972 (0.0734)	0.8000 (0.2976)	0.8000 0.0783
$\hat{b}_M$ (by (3.5))	0.8010 (0.0915)	0.8005 (0.0363)	1.1004 (0.6265)	0.8005 (0.3306)
$\hat{b}_U$ ($a = 0.5$)	0.7940 (0.0982)	0.7981 (0.0742)	0.8000 (0.0000)	0.8000 (0.0014)
$\hat{b}_U$ ($a = 2.0$)	0.7968 (0.0975)	0.8000 (0.0648)	0.8000 (0.0011)	0.7996 (0.0128)
$\hat{b}_U$ ($a = iqd(y)/\sqrt{n}$)	0.7927 (0.1048)	0.7987 (0.0764)	0.8000 (0.0000)	0.8000 (0.0012)

and

$$\sum_{t=1}^{n} |y_t - \hat{b}_0 - \hat{b}_M x_t| = \inf_{\beta_0, \beta} \sum_{t=1}^{n} |y_t - \beta_0 - \beta x_t|. \tag{3.5}$$

Table I illustrates the robustness of $\hat{b}_U$ in relation to $\hat{b}_L$ and $\hat{b}_M$. The estimators $\hat{b}_L$ and $\hat{b}_M$ perform poorly in some cases while $\hat{b}_U$ is acceptable in all cases. For normal noise $\hat{b}_L$ by (3.1) is, not surprisingly, the best. The accuracy of estimator $\hat{b}_U$ for large a is similar to that for $\hat{b}_M$. For Cauchy noise $\hat{b}_L$ has an extremely large spread in comparison to both $\hat{b}_U$ and $\hat{b}_M$. This is due to the fact that the noise has infinite variance. For Bernoulli noise $\hat{b}_M$ has a large spread, and $\hat{b}_U$ is clearly the best. The estimator $\hat{b}_U$ is also the best in the case of asymmetric noise.

The choice of the adjustable parameter a is a problem that deserves some attention. It is encouraging to note that the estimator $\hat{b}$ is not too sensitive to the choice of a, according to the simulations shown in Table I and Fig. 1. In practice it seems reasonable to choose a somehow proportional to y_t. This insures that the estimated model is invariant to scale transformations of y.

TABLE II

Simulations for Model (3.6) with $b = 0.8$

Distribution of ε_t	$U(0, 9)$	$C(0, 1)$	*BERN*	*ASYM*
$\hat{b}_L$ (by (3.3))	0.7848 (0.0818)	0.7934 (0.0449)	0.7892 (0.0811)	0.7981 (0.0556)
$\hat{b}_M$ (by (3.5))	0.7829 (0.1318)	0.7991 (0.0079)	0.7588 (0.5816)	0.8010 (0.039)
$\hat{b}_U$ ($a = 0.5$)	0.7733 (0.0862)	0.7989 (0.0108)	0.8000 (0.0000)	0.8000 (0.0003)
$\hat{b}_U$ ($a = 2.0$)	0.7674 (0.0875)	0.7992 (0.0080)	0.7997 (0.0008)	0.7999 (0.0010)
$\hat{b}_U$ ($a = iqu(y)/\sqrt{n}$)	0.7725 (0.0879)	0.7986 (0.0088)	0.8000 (0.0000)	0.8000 (0.0009)

Table II reports the same simulation results for the following autoregressive model with $b = 0.8$:

$$y_t = by_{t-1} + \varepsilon_t, \qquad t = 1, ..., n. \tag{3.6}$$

The results are similar to that of the previous regression model.

On the other hand, for the case where x has higher dimension, computational complexity should be involved. Actually, the calculation of $A_n(\beta)$ involves $O(n^2)$ operations. Therefore $\hat{b}_U$ is computationally more costly than $\hat{b}_L$ or $\hat{b}_M$. The following algorithm may be a promising one:

- Step 0. Obtain an initial estimator $\hat{b}_1$. For instance, the LSE may be as the initial one.
- Step 1. Use the conjugate gradient method to update $\hat{b}_1$.
- Step 2. Continue Step 1 until convergence to obtain the final estimator $\hat{b}_U$.

4. STRONG CONSISTENCY

In this section, we prove that the new estimator $\hat{b}_U$ defined by (2.11) is strongly consistent to the true value b in the two cases mentioned in Section 1. We first consider the autoregressive case.

THEOREM 4.1. *Let $\hat{b}_U$ be the estimator of b for model* (1.4), *and let it be determined by* (2.11). *Under assumption B defined in Section* 1, *we have*

$$\lim_{n \to \infty} \hat{b}_U = b \qquad a.s. \tag{4.1}$$

In order to prove this theorem, first we introduce some notations and then establish a lemma. Let

$$\eta^\tau = (-x^\tau, \varepsilon), \qquad \eta_t^\tau = (-x_t^\tau, \varepsilon_t).$$

$F_\eta(\cdot)$ stands for the distribution of η. $\hat{F}_\eta(\cdot)$ is the empirical distribution based on the sample $\{\eta_t^\tau : t = 2, ..., n\}$

LEMMA 4.2. *Suppose that $\{\eta_t^\tau : t = 1, ..., n\}$ is a stationary and ergodic sequence with common distribution $F_\eta(\cdot)$. Then*

$$\lim_{n \to \infty} \sup_{u \in R^{p+1}} |\hat{F}_\eta(u) - F_\eta(u)| = 0 \qquad a.s. \tag{4.2}$$

Proof. Define a class of indicator functions by

$$\mathscr{F}_0 = \{ f_u(w) : f_u(w) = I(-\infty < w \leqslant u), u \in R^{p+1} \},$$

where $w = (w_1, ..., w_{p+1})^\tau$, $u = (u_1, ..., u_{p+1})^\tau$, and $I(-\infty < w \leqslant u) = 1$ if $w_i \in (-\infty, u_i]$ for $i = 1, ..., p+1$; and equals 0 otherwise. We use f as an abbreviation for $f_u(w)$ below. It is easy to see that for any fixed $\delta > 0$ there exists a finite class of indicator functions $\mathscr{F}_\delta$ such that for any fixed $f \in \mathscr{F}_l$,

$$f_{\delta l} \leqslant f \leqslant f_{\delta u} \qquad \text{and} \qquad P(f_{\delta u} - f_{\delta l}) \leqslant \varepsilon$$

for some $f_{\delta l}$ and $f_{\delta u} \in \mathscr{F}_\delta$, where P denotes the probability measure corresponding to the distribution $F(\cdot)$, and

$$Pf =: \int f \, dP.$$

These notations are used by Pollard (1984). The proof is completed by making use of Theorem II.2 with the supplementary remark of Pollard (1984, pp. 8–9).

COROLLARY 4.3. *Let $\varphi_\eta(v)$ and $\hat{\varphi}_\eta(v)$ be the characteristic functions of $F_\eta(\cdot)$ and $\hat{F}_\eta(\cdot)$ respectively. Under the conditions required by Lemma* 3.1, *for any $c > 0$,*

$$\lim_{n \to \infty} \sup_{\|v\| \leqslant c} |\hat{\varphi}_\eta(v) - \varphi_\eta(v)| = 0 \qquad a.s., \tag{4.3}$$

where $\|\cdot\|$ denotes the Euclidean norm on R^{p+1}.

Proof. By making use of properties of the characteristic function (e.g., see Theorem 8.3.3 and Corollary 8.3.5 of Chow and Teicher, 1978), (4.3) follows from (4.2) directly.

The Proof of Theorem 4.1. Let $\xi^\tau = (b - \beta)^\tau$. Recalling the definition of the above $\varphi_\eta(\cdot)$ and $\varphi(s, \beta)$ in (2.2), we have

$$\varphi_\eta(\xi^\tau s, s) = Ee^{is(\xi^\tau x + \varepsilon)} = \varphi(s, \beta)$$

and then by (2.3)

$$A(\beta) = \int |\varphi_\eta(\xi^\tau s, s)|^2 \, w(s) \, ds.$$

Similarly, by (2.9) and (2.11),

$$A_n(\beta) = \int |\hat{\varphi}_\eta(\xi^\tau s, s)|^2 \, w(s) \, ds.$$

For any fixed $\delta > 0$, there exists a positive c such that

$$\int_{-\infty}^{-c} w(t) \, dt + \int_{c}^{\infty} w(t) \, dt < \delta.$$

Note that by the definition (2.12) D_p is bounded, thus by the above inequality and (4.3) we have

$$\begin{aligned} &\limsup_{n \to \infty} \sup_{\beta \in D_p} |A_n(\beta) - A(\beta)| \\ &\quad \leqslant \limsup_{n \to \infty} \int_{-c}^{c} \sup_{\beta \in D_p} |\hat{\varphi}_\eta((b - \beta)^\tau s, s) - \varphi_\eta((b - \beta)^\tau s, s)| \, w(s) \, ds \\ &\quad + 2 \int_{-\infty}^{-c} w(t) \, dt + 2 \int_{c}^{\infty} w(t) \, dt \leqslant 2\delta \qquad \text{a.s.} \end{aligned}$$

Thus, since δ is arbitrary in this equation, we have

$$\lim_{n \to \infty} \sup_{\beta \in D_p} |A_n(\beta) - A(\beta)| = 0 \qquad \text{a.s.}$$

Note that $A_n(\beta)$ and $A(\beta)$ are continuous, and they attain their maximum values at $\hat{b}_U$ and b respectively, and by (2.7), $A(b) > A(\beta)$ for $\beta \neq b$. Thus it is easy to see that (4.1) follows from (4.3). The proof of Theorem 4.1 is completed.

Now we consider the i.i.d. case.

THEOREM 4.4. *Suppose that* $\{(y_t, x_t^\tau): t = 1, ..., n\}$ *satisfies model* (1.2) *with condition* (*A*), $\{(y_t, x_t^\tau): t = 1, ..., n\}$ *are i.i.d., and* $\hat{b}_U$ *is determined by* (2.11). *Let the density kernel* $w(t)$ *chosen in* (2.3) *satisfy* (2.4) *and*

$$\int \left| \frac{d\varphi_w(x)}{dx} \right| dx < \infty \tag{4.4}$$

where $\varphi_w(x)$ *is defined in* (2.9). *Then*

$$\lim_{n \to \infty} \hat{b}_U = b \qquad a.s.$$

Proof. Because $\{y_t, x_t^\tau)$, $t = 1, ..., n$ are i.i.d., $\{(y_t, x_t^\tau)\}$ is stationary and ergodic, and then the results of Lemma 4.2 and Corollary 4.3 hold again. But the proof of Theorem 4.1 fails to work because D_p is replaced by R^p here, which is not bounded. Thus we add the restriction (4.4) for density kernel $w(t)$. Now we put

$$F_\beta^*(x) = \int F_\beta(x+u)\, dF_\beta(u), \qquad \hat{F}_\beta^*(x) = \int \hat{F}_\beta(x+u)\, d\hat{F}_\beta(u), \tag{4.5}$$

and then by (2.8) and (2.10)

$$A(\beta) = \int \varphi_w(x)\, dF_\beta^*(x), \qquad A_n(\beta) = \int \varphi_w(x)\, d\hat{F}_\beta^*(x). \tag{4.6}$$

Because of condition (4.4), it is possible to use integration by parts to rewrite $A(\beta)$ as

$$A(\beta) = \int F_\beta^*(x)\, \psi_w(x)\, dx, \qquad A_n(\beta) = \int \hat{F}_\beta^*(x)\, \psi_w(x)\, dx,$$

where

$$\psi_w(x) = d\varphi_w(x)/dx.$$

By making use of Lemma 4.1 and (4.5), it follows that

$$\lim_{n\to\infty} \sup_{x,\beta} |\hat{F}^*_\beta(x) - F^*_\beta(x)| = 0 \qquad \text{a.s.}$$

Consequently, combining with (4.6) we have

$$\lim_{n\to\infty} \sup_{\beta} |A_n(\beta) - A(\beta)| \leqslant \lim_{n\to\infty} \int |\hat{F}_\beta(x) - F^*_\beta(x)|\, |\psi_w(x)|\, dx = 0 \qquad \text{a.s.}$$

The remainder of the proof is the same as the description below (3.9) in the proof of Theorem 4.1. The proof is completed.

Remark 4.5. The restriction (4.4) is not strong, because many density kernel functions satisfy (3.10), e.g., the densities of $N(0, \sigma^2)$ and of the symmetric exponential distribution (see (2.14)). In particular, $\varphi_w(x)$ of (2.15), which was used in the simulation results, also satisfies (4.4).

5. FURTHER DISCUSSION

First, we discuss the asymptotic normality of the estimator $\hat{b}_U$. It is obvious that we cannot provide central limit results under only assumption A, which is too weak. At least some moment conditions on x_t, for model (1.2), are needed. On the other hand, if the asymptotic normality of $\sqrt{n}\,(\hat{b}_U - b)$ is obtained, it will depend not only on the distributions of x_t and ε_t of (1.2), but also on the density function $w(\cdot)$ of (2.4). We can show the asymptotic normality of $\hat{b}_U$ by using the results of the so-called "U-Processes" (cf. Nolan and Pollard, 1988) and the idea of Theorem VII.1.5 of Pollard (1984) under certain conditions for the i.i.d. case, i.e., $\{(y_t, x_t^\tau)\}$ is an i.i.d. series, but we have not yet done that for the autoregressive case (see model (1.3)). We now describe briefly how to show the asymptotic normality.

For the i.i.d. case, let (ε, x^τ) be a random vector, $(\varepsilon', (x')^\tau)$ be an independent copy of (ε, x^τ), and $\zeta^\tau = (\varepsilon - \varepsilon', (x - x')^\tau)$, $\theta^\tau = (b - \beta)^\tau$. Suppose that $(\varepsilon_t, x_t^\tau)$, $t = 1, ..., n$, are the independent n-observations of (ε, x^τ). We regard

$$\zeta_{t,s} = (\varepsilon_t - \varepsilon_j, (x_t - x_j)^\tau), \qquad t, j = 1, 2, ..., n,$$

as the n^2-observations of ζ, which are not independent, of course. Let P_ζ denote the probability measure generated by random vector ζ.

Thus by (2.8), it is easy to see that

$$A(\beta) = \int \varphi_w((1, \theta^\tau)\,\zeta)\, dP_\zeta,$$

where $\varphi_w(\cdot)$ is determined by (2.9). We introduce the following notations which are similar to those used by Pollard (1984):

$$\varphi_w(\zeta, \theta) = \varphi_w((1, \theta^\tau)\,\zeta),$$

and $\varphi_w(\cdot, \theta)$ is the abbreviation of $\varphi_w(\zeta, \theta)$, and

$$P\varphi_w(\cdot, \theta) = \int \varphi_w((1, \theta^\tau)\,\zeta)\, dP_\zeta \equiv F_w(\theta).$$

Define 0 as a zero p-vector. Suppose that $\varphi_w(\cdot, \theta)$ has a linear approximation near the 0, at which $F_w(\theta)$ attains its maximum value (see (2.7))

$$\varphi_w(\cdot, \theta) = \varphi_w(\cdot, 0) + \theta^\tau \Delta(\cdot) + \|\theta\|\, r(\cdot, \theta). \tag{5.1}$$

Let P_n be the empirical distribution of ζ based on the samples $\zeta_{t,s}$, $t, s = 1, ..., n$, i.e.,

$$P_n(\zeta = \zeta_{t,s}) = \begin{cases} 1/n^2 & \text{for} \quad t, s = 1, 2, ..., n, \\ 0 & \text{otherwise.} \end{cases}$$

Recalling the definition of $A_n(\beta)$ in (2.10), we have

$$A_n(\beta) = P_n\varphi(\cdot, \theta) \equiv F_n(\theta).$$

Under assumption A with some other restrictions, for example, that $E(xx^\tau)$ is finite, we could obtain that, via the idea of Theorem VII.1.5 of Pollard (1984),

$$\sqrt{n}\,(\hat{b}_U - b) \Rightarrow N(0, 4V^{-1}(P\Delta_1\Delta_1^\tau)\,V^{-1})$$

where $\Delta_1 = \Delta_1(\varepsilon_1, x_1) = E(\Delta_1 \mid \varepsilon_1, x_1)$ means the conditional expectation of Δ given (ε_1, x_1), V is the second derivative matrix of $F_w(\theta)$ at θ_0, and $(P\Delta_1\Delta_1^\tau) = \int \Delta_1\Delta_1^\tau\, dP_\zeta$ is the covariance matrix of $\Delta(\cdot)$ in (4.5) (noting $P\Delta = 0$). By some calculation, we know that

$$V^{-1}[P\Delta_1\Delta_1^\tau]\,V^{-1} = C_w\Gamma^{-1},$$

where $\Gamma = E(x - Ex)(x - Ex)^\tau$, and C_w is a positive number depending on $w(\cdot)$ of (2.4) and on the distribution of ε_t as well. As an example, consider

a $w(t)$ defined by (2.13). In this case, in addition to assumption A, only two more conditions are needed, i.e., Γ is finite and $E(3\varepsilon^2 - a^2)/(a^2 + \varepsilon^2)^3 \neq 0$, to ensure (4.7) with

$$C_w = E\left\{E\left[\frac{\varepsilon_1 - \varepsilon_2}{(a^2 + (\varepsilon_1 - \varepsilon_2)^2)^2} \,\middle|\, \varepsilon_1\right]\right\}^2 \left\{E\left\{\frac{3\varepsilon^2 - a^2}{(a^2 + \varepsilon^2)^3}\right\}\right\}^{-2},$$

where $E(\cdot \mid \varepsilon_1)$ stands for the conditional expectation given ε_1. The general case may be complicated but without any essential difficulty.

Pollard's result treats the empirical processes, not U-processes. But a similar conclusion can be reached by means of Pollard's idea. In order to see clearly how to employ the idea of Pollard's conclusion (1984, Theorem VII.1.5.), we now cite his theorem as follows.

THEOREM 5.1. *Suppose $\{s_n\}$ is a sequence of random vectors converging in probability to the value t_0 at which $F(\cdot) = \int f(\cdot, x)\, dP(x)$ has its minimum (in our case, the maximum is considered). Define $r(\cdot, s)$ and the vector function $\Delta(\cdot)$ by*

$$f(\cdot, s) = f(\cdot, s_0) + (s - s_0)^\tau \Delta(\cdot) + \|s - s_0\|\, r(\cdot, s).$$

If

(i) *s_0 is an interior point of the parameter set T;*

(ii) *$F(\cdot)$ has a nonsingular second derivative matrix V at s_0;*

(iii) *$f_n(s_n) = o_p(n^1) + \inf_s F(s)$;*

(iv) *the components of Δ all belong to $\mathscr{L}^2(P)$; and*

(v) *the sequence $\{\sqrt{n} \int r(x, s)\, d(P_n(x) - P(x)\}$ is stochastically equicontinuous at s_0,*

then

$$\sqrt{n}\,(s_n - s_0) \Rightarrow N(0,\, V^{-1}(P(\Delta\Delta^\tau) - (P\Delta)(P\Delta^\tau))\, V^{-1}),$$

where $F_n(\cdot) = \int f(\cdot, x)\, dP_n(x)$, $P\Delta\Delta^\tau = \int \Delta(x)\, \Delta^\tau(x)\, dP(x)$, and P_n is the empirical measure based on the sample.

In our case, the stochastic process investigated is a U-process. Roughly speaking, a U-process is a set of U-statistics. In order to obtain asymptotic normality we only need to check, in view of the proof of the theorem, whether or not similar conditions in Pollard's theorem are fulfilled in the U-process case. Note that the condition (v) in Pollard's theorem was stated in terms of stochastic equicontinuity in the empirical process case (see

Pollard, 1984, Lemma VII.15, pp. 150). Meanwhile, such a stochastic equicontinuity in the U-process case continues to hold as well (cf. Nolan and Pollard, 1988). Hence, the analogy of Condition v, in our case, can be checked. The other conditions are easily satisfied.

Looking at the proof of the theorem (see Pollard, 1984, pp. 141–142), all the work that remains is checking the asymptotic normality of $\sqrt{n}\int \Delta(x)\, d(P_n - P)$ in our case. This is done by a well-known result because it is nothing but a U-statistic. Consequently, the asymptotical normality can be verified.

Now we have a brief observation for the robustness of the estimator $\hat{b}_U$, which is denoted by $\hat{b}_{Un}$ below and maximizes $A_n(\beta)$ in (2.11). One can easily check that if any one of the values $\varepsilon_1, ..., \varepsilon_n$ tends to infinity, say $|\varepsilon_n| \to \infty$, then $\hat{b}_{Un}$ tends to $\hat{b}_{U(n-1)}$ which is the same estimator of b, but is based on $(y_1, x_1^\tau), ..., (y_{n-1}, x_{n-1}^\tau)$ only. This fact implies that $\hat{b}_U$ is robust in some sense. Indeed, it is easy to see that the gross-error sensitivity of b at the distribution P of (y, x^τ) in Hampel's sense is finite. Consequently, b is B-robust at P in the sense of Rousseeuw (1981). The simulation results shown in Tables I and II in Section 2 also demonstrate this fact. But if we try to describe the robustness $\hat{b}_U$ with the concept of a breakdown point (e.g., Hampel *et al.*, 1986), we meet some difficulties. Thus we leave this subject to future research.

Another consideration is to use the procedure of Section 2 for estimating regression coefficients of the following partial regression model, i.e. (e.g., see Schick, 1986),

$$y = b^\tau x + g(z) + \varepsilon,$$

where x, z, ε are independent of each other, and $g(\cdot)$ is an unknown smooth function. Let $\varepsilon' = g(z) + \varepsilon$, then the model above becomes

$$y = b^\tau x + \varepsilon'.$$

It is obvious that only if $(y_t, x_t, \varepsilon_t')$ satisfies assumption A of Section 1 can we obtain a strongly consistent estimator of b, by the procedure of Section 2. When $g(\cdot)$ is required to be estimated, one can consider estimation of $g(\cdot)$ by analyzing the residuals $y_t - \hat{b}_U^\tau x_t$, $t = 1, ..., n$. But in order to ensure some asymptotic properties for the estimation of $g(\cdot)$, some more conditions may be needed.

ACKNOWLEDGMENT

The authors thank the editors and a referee for their comments and suggestions which improved greatly on the early draft of the paper. They would like to thank K. T. Fang as well for helpful suggestions.

REFERENCES

[1] Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Prob.* **12** 1041–1067.
[2] An, H. Z. (1990). Non-negative autoregressive models. Preprint.
[3] An, H. Z., and Chen, Z. G. (1983). On convergence of LAD autoregression with infinite variance. *J. Multivariate Anal.* **12** 335–345.
[4] An, H. Z., Chen, Z. G., and Hannan, E. J. (1982). Autocorrelation, autoregression and autoregressive approximation. *Ann. Statist.* **10** 926–936.
[5] Anděl, J. (1989). Non-negative autoregressive processes. *J. Time Ser. Anal.* **10** 1–11.
[6] Bell, C. B., and Smith, E. P. (1986). Inference for non-negative autoregressive schemes. *Comm. Statist. Theory Methods* **15** 2267–2293.
[7] Chen, X. R., and Wu, Y. H. (1988). Strong consistency of M-estimates in linear models. *J. Multivariate Anal.* **27** 116–130.
[8] Chow, Y. S., and Teicher, H. (1978). *Probability Theory*. Springer-Verlag, New York.
[9] Hampel, F. R. *et al.* (1986). *Robust Statistics*. Wiley, New York.
[10] Koutrouvelis, I. A. (1982). Regression with stable errors: an empirical characteristic function approach. *Statistica* **42** 209–222.
[11] Lai, T. L., and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10** 154–166.
[12] Nolan, D., and Pollard, D. (1988). Functional limit theorems for U-processes. *Ann. Probab.* **16** 1291–1298.
[13] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.
[14] Reiersol, O. (1950). Identifiability of a linear relation between variable which are subject to error. *Econometrica* **18** 375–389.
[15] Rousseeuw, P. J. (1981). A new infintesimal approach to robust estimation. *Z. Wahrsch. Geb.* **56** 127–132.
[16] Schick, A. (1986). On asymptotically efficient estimation in semiparametric models. *Ann. Statist.* **14** 1139–1151.